Single nucleotide polymorphism (SNP) discovery in porcine expressed genes

S. C. Fahrenkrug, B. A. Freking, T. P. L. Smith, G. A. Rohrer and J. W. Keele

USDA, ARS, US Meat Animal Research Center, Spur 18D, Clay Center, NE, USA

Summary

High-throughput genotyping of swine populations is a potentially efficient method for establishing animal lineage and identification of loci important to animal health and efficient pork production. Markers were developed based upon single nucleotide polymorphisms (SNPs), which are abundant and amenable to automated genotyping platforms. The focus of this research was SNP discovery in expressed porcine genes providing markers to develop the porcine/human comparative map. Locus specific amplification (LSA) and comparative sequencing were used to generate PCR products and allelic information from parents of a swine reference family. Discovery of 1650 SNPs in 403 amplicons and strategies for optimizing LSA-based SNP discovery using alternative methods of PCR primer design, data analysis, and germplasm selection that are applicable to other populations and species are described. These data were the first large-scale assessment of frequency and distribution of porcine SNPs.

Keywords EST, porcine, SNP, swine.

Introduction

Efforts are underway to identify genomic loci suitable for exploitation in livestock genetic improvement programmes. Although quantitative trait loci (QTL) influencing livestock production traits have been reported, their localization is not precise and information about corresponding genomic segments is sparse. Current resolution of QTL regions span from 1 to 50 cM, corresponding to chromosomal regions containing dozens to hundreds of candidate genes. Improved resolution of these regions and identification of candidate genes would be accelerated after establishing comprehensive livestock/human comparative maps. In addition, ongoing and future scans for QTL would be greatly facilitated by the development of linkage maps containing markers amenable to high-throughput genotyping.

Address for correspondence

Brad A. Freking, USDA, ARS, US Meat Animal Research Center, PO Box 166, Spur 18D, Clay Center, NE 68933-0166, USA.
E-mail: freking@email.marc.usda.gov
Present address: S.C. Fahrenkrug, Department of Animal Science, University of Minnesota, 495 An Sci/Vet Med, 1988 Fitch Ave., St Paul, MN 55108, USA.

Accepted for publication 27 November 2001

Current animal genotyping technology relies on the use of simple sequence repeats (SSRs). Although SSRs have been instrumental in generating livestock genetic maps, they are less amenable to high-throughput genotyping instrumentation. In addition, SSR alleles are not always identical-by-decent (IBD), complicating across-family analyses. Unlike SSRs, single nucleotide polymorphisms (SNPs) are abundant, with an average of at least one heterozygous position per 1000 bp in humans (Chakravarti 1999), and amenable to assessment using high-throughput genotyping technologies. In addition, SNPs are extremely stable, occurring at a rate of only one mutation in 2×10^8 nucleotides in humans (Sachidanandam *et al.* 2001). Because the rate of mutation is rare, SNP alleles are almost exclusively IBD and therefore, can be used for QTL scans that extend across pedigrees of distantly related animals. This application will enhance the power of QTL detection, as well as reveal specific populations and individuals suitable for fine-mapping efforts aimed at identifying economically important genetic variation. With the goal of developing high throughput genotyping technologies, we sought to identify SNPs and develop SNP-based marker systems for genetic linkage analysis in pigs.

Single nucleotide polymorphism discovery efforts for humans and mice have relied primarily on two approaches:

(1) accumulation and survey of redundant genomic sequence by whole-genome (Venter *et al.* 2001; IHGSC 2001) and reduced representation shotgun sequencing (RRS, Altshuler *et al.* 2000; Mullikin *et al.* 2000), and (2) locus specific amplification (LSA) and comparative re-sequencing from multiple individuals (Rieder *et al.* 1998). The degree to which variation is sampled with clone-based approaches depends strongly on the degree of redundancy with which the library is sequenced, as well as the diversity of the sampled germplasm. Undiscovered variation can have catastrophic effects on performance of genotyping assays in the form of null alleles. The degree of redundancy required to sufficiently describe extant variation to avoid interference with genotyping assays cannot be estimated without first establishing the diversity of the germplasm under investigation, a parameter heretofore unknown for swine. An effective method for characterizing extant variation at a specific locus for a given population is LSA and comparative sequencing. This method requires the design and synthesis of oligonucleotide primers for each locus, and subsequent amplification and bi-allelic sequence comparison from several individuals (Nickerson *et al.* 1997, 1998; Rieder *et al.* 1998). Heterozygous positions are detected as multiple peaks at the same position in DNA sequence chromatograms.

An LSA-based SNP discovery effort requires a collection of sequences against which PCR amplification primers can be designed. Because of the paucity of porcine sequence data, we initiated a porcine LSA-based SNP-discovery effort in tandem with ongoing livestock expressed sequence tag (EST) sequencing (Smith *et al.* 2001a; Fahrenkrug *et al.* 2002). Genotyping SNPs in the MARC porcine reference family (Rohrer *et al.* 1994) will allow effective positioning of ESTs on the porcine genetic map and subsequent comparative map refinement (Smith *et al.* 2001b).

In the context of a parallel SNP discovery effort in pigs and cattle, primer pairs were developed for EST-associated LSA and assessed for their utility in both species. Alternative methods of PCR primer design were assessed with regard to amplification efficiency and SNP discovery. Quality and informativeness of each putative polymorphism was assessed and the number of animals included in our SNP discovery panel were structured to efficiently discover polymorphisms suitable for genetic mapping using the MARC porcine reference family. These analyses have resulted in the development of an efficient system for EST-associated SNP discovery, as evidenced by the identification and evaluation of 1650 SNP in 403 porcine EST-associated STS. Frequency, distribution, and composition of observed SNPs in western and Chinese pigs was assessed.

Materials and methods

Clones from normalized porcine and bovine cDNA libraries (Smith *et al.* 2001a; Fahrenkrug *et al.* 2002) were sequenced from their 5′-ends, and sequences were compared with GenBank non-redundant and high throughput genomic sequence databases. To maximize the amount of comparative information provided by our SNP discovery programme, we focused our attention on ESTs likely to be orthologous (bit score ≥ 300, Altschul *et al.* 1990) to human sequences with known physical or genetic map positions (GeneMap ′99, Deloukas *et al.* 1998).

PCR primer design

Three approaches to primer design were implemented. First, 'blind' primers were designed using Primer3 (Rozen & Skaletsky 2000) with the optimal PCR product length set between 200 and 800 bp, disregarding any potential intervening genomic sequence. Next, primers designed to amplify 3′ UTR sequences were developed. In order to generate these primers, clones with 5′-end sequences orthologous to mapped human genes were sequenced from their 3′-end. Optimal PCR primers were designed to generate amplicons between 300 and 700 bp long using Oligo 5.0 (National Biosciences, Plymouth, MN, USA). Finally, primers which potentially flank introns were developed. In order to generate intron-directed primers, bovine and porcine EST sequences were compared by Blastn (Altschul *et al.* 1990) to human genomic sequences in the GenBank HTGS database. Blast matches with a total bit score exceeding 300 were predicted to represent the identification of human/livestock orthologues. Because gene structure is highly conserved among vertebrates (Long *et al.* 1995), sequence match discontinuity of at least 50 bp between blocks of high identity (score 80 and homology 80%) were predicted to represent human intron sequence. Amplification primers were designed using Primer 3 from livestock sequence to amplify across a predicted splice-junction, with a preferred amplicon size around 800 bp (software developed by J.W. Keele *et al.*, unpublished data).

PCR and sequencing

Amplification reactions were done in non-skirted, 96-well polypropylene, V-bottom PCR plates (MJ Research, Incline Village, NV, USA). Reactions were prepared according to manufacturers recommendations (Qiagen, Santa Clarita, CA, USA). Each primer pair was initially tested on DNA from three pigs, three cattle, one sheep, and one hamster. Initial PCR conditions were 94 °C for 15 min, 45 cycles of 94 °C for 20 s, 58 °C for 30 s, 72 °C for 1 min, and a final extension at 72 °C for 10 min The result of the PCR reaction

was analysed by 3% agarose gel electrophoresis at 180 mV for 30 min. Results for each individual animal and species were coded and entered into the MARC relational database (MARCDB) as 0 = no product, 1 = primer-dimer, 2 = smear, 3 = multiple bands, 4 = single-light band or 5 = ready to sequence. Conditions for PCR were optimized as required by adjusting annealing temperature. Reactions exhibiting a single, strong band greater than 200 bp in pigs were selected for amplification and sequencing. Primer pairs meeting this criteria were used to amplify a panel of eight or 16 pigs for sequencing and SNP discovery. Products of the PCR reactions (20 μl) were incubated with one unit of Exonuclease I at 37 °C for 1 h, followed by heat inactivation at 75 °C for 20 min. Sequencing of purified PCR products was conducted using ABI Big Dye terminator chemistry and analyzed on an ABI 3700 sequencer.

Interactive polymorphism discovery and tagging

Chromatograms were imported into MARCDB, bases called with Phred (Ewing & Green 1998; Ewing *et al.* 1998) and sequences assembled into contigs with Phrap (P. Green, unpublished data). Polymorphisms were identified using Polyphred (Nickerson *et al.* 1997), interactively assessed using Consed (Gordon *et al.* 1998), and tagged as accepted or rejected. Single nucleotide polymorphisms not detected by Polyphred, and insertion/deletions (indels) were also tagged. Indels were tagged by assigning the inserted allele sequence as consensus and tagging bases extending from one base proximal to one base distal of the indel. Deriving sequence of the deletion or insertion from heterozygous animals in many cases required manual sequence deconvolution. Animal sequences were genotyped as homozygous insertion, deletion or heterozygous insertion/deletion. Sequence replacements of unequal length were tagged with the minimum number of indels to fully describe the replacement. Position and composition of each accepted polymorphism, animal genotypes and contig sequences were parsed to MARCDB. Consensus sequences were submitted to dbSTS (GenBank G72426-G73084). These consensus sequences contain the identified polymorphisms in the form of IUB codes at the consensus position. The IUB codes represent heterozygous positions as follows: R = A/G, Y = C/T, S = G/C, W = A/T, K = G/T, M = A/C, B = C/G/T, D = A/G/T, H = A/C/T, V = A/C/G. Polymorphisms associated with insertion/deletion events were submitted to dbSNP.

Validity and confidence score determination

Allele information associated with each polymorphism was queried from MARCDB using Brio5.0 (Palo Alto, CA, USA). Discrepancies of conflicting genotypes between bi-directional sequences within animal were manually resolved when possible. Genotypic data were then used to generate a measure of polymorphism quality referred to as a validity score (VS) which is defined as

$$\mathrm{VS} = \left[3 \times \left(\frac{1}{1+\mathrm{AA}} + \frac{1}{1+\mathrm{AB}} + \frac{1}{1+\mathrm{BB}}\right)^{-1}\right] - 1$$

where and A and B represent alleles of a biallelic marker, while AA, BB and AB refer to the number of homozygote or heterozygote genotypes. The value for VS is a measure of the number of replications of all three genotypes. This is based on the expectation that the observation of all three genotypes increases confidence that a true SNP exists. To estimate the relationship between VS and expected rate of SNP validation, we established a confidence score (CS) as a probability value where $\mathrm{CS} = 100\% \times [1 - (1 - p)]^{\mathrm{VS}}$ and P is the expected probability that an SNP with a VS = 1 would be validated. The CS value predicts probability of SNP validation given that at least one observation per genotype is identified. Based on our SNP discovery data, we estimated $P = 80\%$ with VS = 1. Therefore VS = 1 (equivalent to a single observation of each homozygote, and a heterozygote) yields CS = 80%, VS = 2 yields CS = 96% and VS = 3 yields CS = 99%.

Within an individual amplicon, amount of potential information for mapping can vary between SNPs, dependent on haplotypes present in the mapping family. We calculated a meiotic index (MX) to prioritize amplicons and SNPs, facilitating development of genotyping assays that maximize the number of unique meioses from the reference family assayed from each amplicon, while minimizing the number of genotyping assays developed. To generate a MX, the CS of each SNP and the meiotic contribution of heterozygous parents to the MARC reference family were considered. The MX for a given amplicon is the cumulative product of the CS and the unique meioses provided by SNP within it.

Results and discussion

Locus specific amplification

Primers for PCR were designed to amplify sequence-tagged sites (STS) corresponding to matched EST using three design strategies (Fig. 1); 'blind', ''3'-end-directed'', and 'intron-directed'. To maximize information from each primer pair, their effectiveness at amplifying a PCR product was assessed using genomic DNA from pigs, cattle, sheep and hamster (to assess potential use in radiation hybrid mapping). Performance of primers designed using each of these approaches is presented in Table 1. Approximately 70% of 157 'blind primer' pairs designed against pig EST sequences

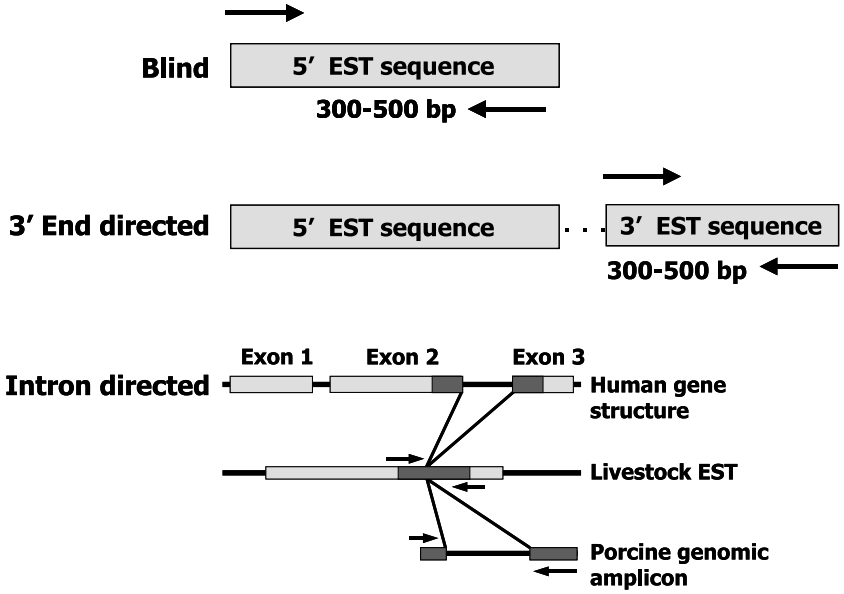

Figure 1 Three different strategies employed for primer design are indicated. Blind primers (PBP, BBP) relied on 5′ EST sequence for development of optimal primers, without consideration of potential splice-junctions. To develop 3′-end directed primers (P3P), 3′-end EST sequence was generated for clones that matched human sequences. For intron-directed primer-pairs (PIP, BIP), livestock ESTs were compared with human genomic sequence and primers developed to amplify across predicted splice-junctions.

(PBP) effectively amplified a PCR product from pig genomic DNA, with an average amplicon size of 437 bp. Cross-species amplification revealed that 46% of PBP generated a PCR product from cattle DNA, 25% from sheep DNA and 29% from hamster DNA. Based on disparity in observed and predicted amplicon sizes for pig, an estimate of frequency of intron-capture was determined for successful amplification products (when observed size exceeded predicted size by 75 bp). Based upon this criterion, 43% of those primers that amplified porcine genomic DNA resulted in serendipitous intron capture.

Primers designed against the 3′-ends of porcine cDNAs (111 pair, P3P) successfully amplified pig genomic DNA 88% of the time. Using the same criterion as above, only 26% of P3P-derived amplicons were predicted to contain an intron. This decreased intron yield was anticipated given that 3′-end sequences correspond to terminal exons, which average 600–700 bp of uninterrupted sequence (Krizman & Berget 1993). The P3P primers resulted in successful PCR product generation from cattle (44%), sheep (38%) and hamster (46%).

It was postulated that primers which amplified non-coding intragenic sequence, such as untranslated regions and introns, yielded highly polymorphic amplicons because non-coding DNA is subject to less selection pressure than coding DNA (Chakravarti 1999). Primers for PCR were therefore designed to amplify across predicted splice-junctions based on alignment of livestock ESTs and human genomic sequence. Intron-directed primers based on pig ESTs (PIP) generated a PCR amplification product suitable for sequencing from pig DNA 78% of the time, with an average amplicon size of 558 bp and a 64% intron-capture rate. Forty-seven percent of PIP pairs also amplified cattle DNA, 41% sheep DNA and 29% hamster DNA. Cross-species intron-capture efficiencies ranged from 55% for hamster, to 64% for sheep.

Intron-directed primers designed for a parallel cattle effort (W.M. Grosse *et al.*, unpublished data) against bovine ESTs (BIP) were also assessed for their cross-species performance. A subset of BIP where data indicated amplification of a large PCR product from porcine genomic DNA were selected for pig SNP discovery. This pre-selection process resulted in 87% of selected BIP amplifying pig genomic DNA when PCR was repeated. A majority of these primers also amplified bovine (83%) and sheep (71%) DNA. As expected, selection of BIP primers that amplified large ready to sequence PCR products in pigs and cattle also resulted in an increased frequency of intron capture relative to other primer types.

An objective of our ongoing SNP discovery effort is refinement of the porcine genetic comparative map. It is, therefore, important that our efforts identify SNP in targeted genes. Porcine STS sequences were compared with the EST database using Blastn (Altschul *et al.* 1990). The majority (90%) of STS corresponded to the same gene as that

Table 1 Primer success, product size from PCR amplification of genomic DNA, and intron capture rates by target species and primer type.

Target species	Primer type[1]	Number pairs tested	Number pairs amplified (%)	Average size (bp)	Number pairs predicted intron capture[2] (% of those amplified)
Pig	PBP	157	110 (70)	437	47 (43)
	P3P	111	98 (88)	441	25 (26)
	PIP	293	229 (78)	558	147 (64)
	BIP	473	412 (87)	686	358 (87)
Cattle	PBP	157	72 (46)	451	29 (40)
	P3P	111	49 (44)	445	15 (31)
	PIP	293	138 (47)	508	84 (61)
	BIP	328	272 (83)	713	239 (88)
Sheep	PBP	157	39 (25)	424	16 (41)
	P3P	111	42 (38)	405	12 (28)
	PIP	288	118 (41)	548	76 (64)
	BIP	328	233 (71)	704	207 (89)
Hamster	PBP	157	46 (29)	368	22 (48)
	P3P	111	51 (46)	410	11 (22)
	PIP	293	85 (29)	419	47 (55)
	BIP	328	131 (40)	750	117 (89)

[1]PBP = primers designed from porcine 5′ EST sequence without regard for intron/exon boundaries; P3P = primers designed from porcine 3′ EST sequence without regard for intron/exon boundaries; PIP = primers designed from porcine EST sequence directed to flank a predicted intron/exon junction; BIP = primers designed from bovine EST sequence directed to flank a predicted intron/exon junction.
[2]Amplification considered to have captured an intron when observed genomic product size exceeded predicted product size from EST sequence by 75 bp.

predicted from EST sequence. Targeting efficiency of each primer-type is indicated in Table 2. Failure to detect a match for some STS could either occur from amplification of an unintended genomic locus, or failure to detect a sequence match. Failure to detect a match for some STS probably results from a lack of exon sequence caused by the placement of intron-directed primers too close to splice-junctions. Discovery of polymorphisms associated with STS that do not correspond to intended targets will nonetheless serve as a resource for SNP-based type-II markers.

SNP discovery

Primers of all types that successfully amplified a single product from pig DNA were used to amplify genomic DNA from either a preliminary ($n = 16$) or modified ($n = 8$; see

Table 2 Summary of porcine genomic sequence acquisition, polymorphism frequency, and rate of success in obtaining sequence for the intended EST by primer type.

Primer type[1]	# Contigs sequence	Contig bp sequenced[2]	Polymorphic contigs	# SNPs	bp/SNP	Targeting efficiency[3] (%)
PBP and BBP	104	48 759	59 (57%)	205	238	88
GSP	30	15 781	22 (73%)	83	190	na
P3P	79	33 293	56 (71%)	168	198	97
PIP and BIP	343	205 190	266 (78%)	1194	172	87
Total	556	303 030	403 (72%)	1650	184	90

[1]PBP = primers designed from porcine 5′ EST sequence without regard for intron/exon boundaries; P3P = primers designed from porcine 3′ EST sequence without regard for intron/exon boundaries; PIP = primers designed from porcine EST sequence directed to flank a predicted intron/exon junction; BIP = primers designed from bovine EST sequence directed to flank a predicted intron/exon junction; GSP = Gene-specific primers designed from multiple species sources directed to flank introns and terminal exons.
[2]Base pairs of contig sequence represented by at least two animals and a quality PHRED score 20.
[3]Percentage of genomic amplicons that matched the EST against which primers were originally designed.

below) SNP discovery panel (MARCSDP) that initially included nine parents of the MARC porcine reference population (Rohrer *et al.* 1994), and seven animals likely to harbour western or Chinese breed specific alleles present in these mapping animals.

A total of 556 EST-associated STS were successfully amplified and sequenced (high quality consensus >100 bp in length). Total high-quality (PHRED > 20) sequence contributed by animals of the discovery panel was nearly 3 Mb (non-overlapping within individual). The scanned sequence corresponds to 303 030 bp of the porcine genome. Polymorphisms were detected in 403 amplicons (72% overall, 1650 SNPs) with SNP frequencies delineated with regard to primer-type in Table 2. Included in this data set are some amplicons generated using 'blind' primers developed against bovine ESTs (BBP) and gene-specific primers (GSP) developed to target introns and terminal exons. Clearly, the likelihood that an amplicon was polymorphic depended on the rationale used in primer design (Table 2). Only 57% of PBP and BBP contigs resulted in detection of polymorphisms, which may in part be because of a small average amplicon size (Table 1). However, although size of P3P derived amplicons was not much different from PBP, nearly 71% of P3P contigs were polymorphic. It is likely that the increased rate of polymorphism for P3P is also dependent on placement of primers with regard to open reading frames (ORFs) within the targeted EST. Strategies for increasing the amount of non-coding DNA being surveyed, such as for P3P and PIP, would be expected to provide more polymorphisms because of a diminished sensitivity of sequence in introns and 3′-UTRs to evolutionary selection. Overall frequency of SNPs in sequenced contigs was 1 SNP per 184 bp. Frequency of SNP per bp also depended on primer type, with amplicons derived from primers directed against non-coding and intergenic regions having greater SNP density relative to amplicons derived from 'blind' primers.

Composition of polymorphisms was 65% C/T or A/G, 14% G/T or A/C, 7.6% G/C, 3.7% A/T, and 9.3% indels. For the remaining 0.4% of the polymorphisms, three alleles were detected. Indels were detected at a rate of about one of twelve of that for SNPs, consistent with results described for humans (Wang *et al.* 1998) and were not derived exclusively from a specific germplasm or cross. Although indels represented only 9.3% of total polymorphisms, they were detected in 20% of the sequenced amplicons and 27.5% of the polymorphic amplicons, regardless of primer-type.

SNP quality and informativeness

Confidence in an individual polymorphism is bolstered by detection, and uniform representation of heterozygotes and homozygotes for both alleles of a given SNP. These properties are reflected by the calculated VS. The VS associated with our SNP ranges from 0.2 to 11.6, with an average VS of 2.4 (SD = 2.15) and 85% of the SNP with a VS 1 (Fig. 2a). An empirical CS was established based on empirical evidence that VS = 1 indicated a true SNP with a probability of 80%. When the complete data set was evaluated for confidence (Fig. 2b), 92% of SNPs were characterized by a CS greater than 75%, and an average CS of 89% (SD = 10). In order to assess our indexing system for establishing SNP validity/confidence, we examined fidelity with which heterozygosity was detected using a technique other than DNA sequencing. Animals comprising of the MARC pig reference family were genotyped by microsequencing-coupled MALDI-TOF mass spectrometry (Buetow

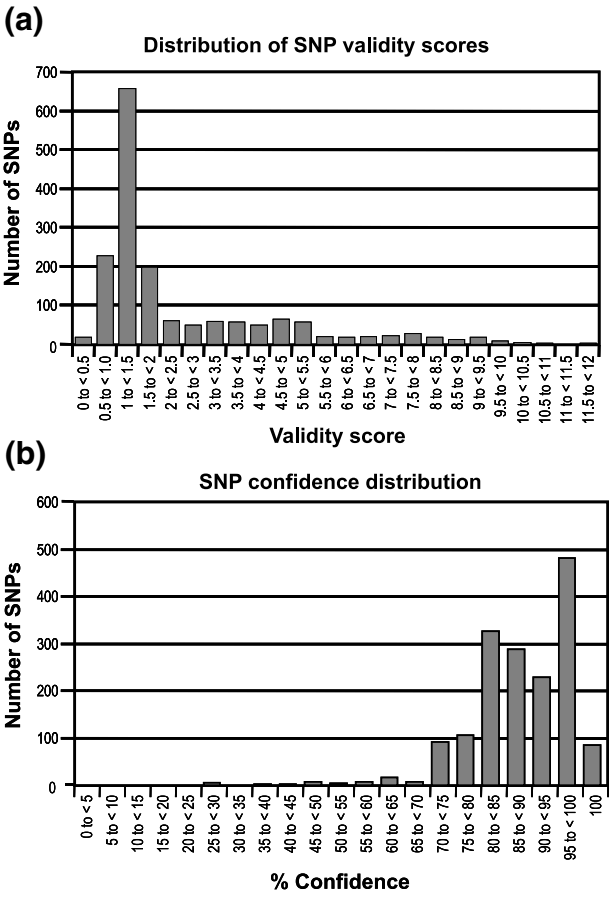

Figure 2 (a) All 1650 predicted polymorphisms were assessed for their 'validity score' (VS). The number of polymorphisms are plotted with respect to the calculated VS range. (b) Confidence values derived for each polymorphism was determined assuming 80% confidence in a polymorphism with a VS = 1. The number of polymorphisms are plotted with respect to the calculated confidence value.

et al. 2001) for 124 SNPs with VS ranging from 0.4 to 9.76 and a mean score of 3.94. Nearly all (97.6%) of these SNPs were verified as polymorphic by MALDI-TOF genotyping, indicating that our confidence estimates closely approximated the observed rate of validation from sequence data alone. Three putative SNPs not verified as polymorphic by MALDI-TOF genotyping had VS scores of 7.82, 1.41, and 1.28. Inspection of the data associated with the SNP with a VS = 7.82 (˜100% confidence) suggested a problem with the genotyping assay. The other two putative SNPs remain unresolved. In summary, our validity and confidence indexing system has proven effective at ranking putative polymorphisms, and will be valuable for maximizing development of informative markers. Using this data to calculate MX for each amplicon, 83% of the amplicons are targets for genotyping assay design and mapping because they are predicted to generate 20 meioses.

SNP discovery germplasm

Relevant questions to address during any polymorphism discovery process include sample size, degree of heterozygosity or diversity in the sample, and extension of the distribution of observed variation to the general population. Sample size is dictated by a balance of throughput capabilities and identification of most of the available allelic variation. The 16 animals initially used in the polymorphism screen (MARCSDP) were selected to represent most of the segregating alleles in the mapping family and to provide seven animals more likely than the F_1 mapping parents to be homozygous at any given SNP. After collection of preliminary data an analysis of the number and quality of SNPs with respect to each animal of the MARCSDP was conducted to determine if the discovery panel could be reduced without substantial loss in detection of informative positions. Cumulative MX contributed by each individual in the MARCSDP for 278 amplicons was assessed. As indicated in Table 3, removal of seven animals of the MARCSDP resulted in only a 2.6% loss in total mapping information, revealing that the presence of purebreds was not significantly contributing to either the quality or MX of our putative SNPs. Thus, these seven animals were removed from MARCSDP. In order to conform to a 96-well plate configuration, and retain Duroc representation, animal 199035702 (Meishan × White composite) was also removed from the panel. This reduced panel resulted in a 7.4% loss in predicted mapping information; however, potential sequencing throughput of number of amplicons was doubled. Estimates of MX after implementation of the reduced panel included only intron-directed primers while estimates of the full panel included all primer types, accounting for similar estimates of total MX, despite a predicted 7.4% loss.

Table 3 Analysis in retrospect of the SNP discovery panel for 278 amplicons summarizing the impact of removing contributing members of the panel on the number of potential meioses discovered in the full panel of 16 animals.

Animal(s) removed	Potential meioses undiscovered[1]	Percentage of total potential meioses from full panel
198893304	21.6	0.1
199031304	1818.5	9.3
199032102	2263.5	11.5
199035602	1387.0	7.1
199035702	780.4	4.0
199035802	771.2	3.9
199035902	2013.4	10.3
199036703	1994.1	10.2
199036705	2019.8	10.3
199043006	6749.2	34.4
199045403	44.4	0.2
199105908	29.2	0.1
199110803	28.8	0.1
199120807	24.0	0.1
199204909	41.7	0.2
199205910	17.2	0.1
Seven lowest	501.2	2.6
Seven lowest and 199035702	1447.2	7.4
Seven lowest and 199035802	1354.9	6.9

[1]Values represent the change in the total meiotic index (MX) for 278 amplicons.

Heterozygosity is a limiting factor for resolution of linkage and linkage-disequilibrium mapping, as well as expected yield from large-scale SNP discovery efforts. Western breed animals and purebred Meishan were found to harbour one heterozygous base for every 895 bp surveyed (SD = 98). Purebred Minzhu and 7/8 Minzhu crossbreds were more polymorphic, with one heterozygous base every 540 bp (SD = 81). The F_1 Chinese-western crossbreds maintained a heterozygous base every 478 bp (SD = 15.5). These values can be used to estimate the number of bases, on average, that must be scanned in order to discover a polymorphism in a specific individual.

Measures of diversity can be estimated by the average heterozygosity per nucleotide (H), and the proportion of sites harbouring variation (K). H does not depend on sample size, whereas K increases with the number of genomes surveyed. When Θ (population parameter for genetic diversity) is small, $H \approx \Theta$ and $K \approx \Theta\,[1^{-1} + 2^{-1} + 3^{-1} + 0 + (n-1)^{-1}]$ (Wang *et al.* 1998). Thus, because Θ and K can be estimated given H (Hartl & Clark 1997), heterozygosity observed in SNP discovery can potentially be used to define these genetic parameters for pigs. Based upon the value for H derived for individuals of the MARCSDP an estimate of K

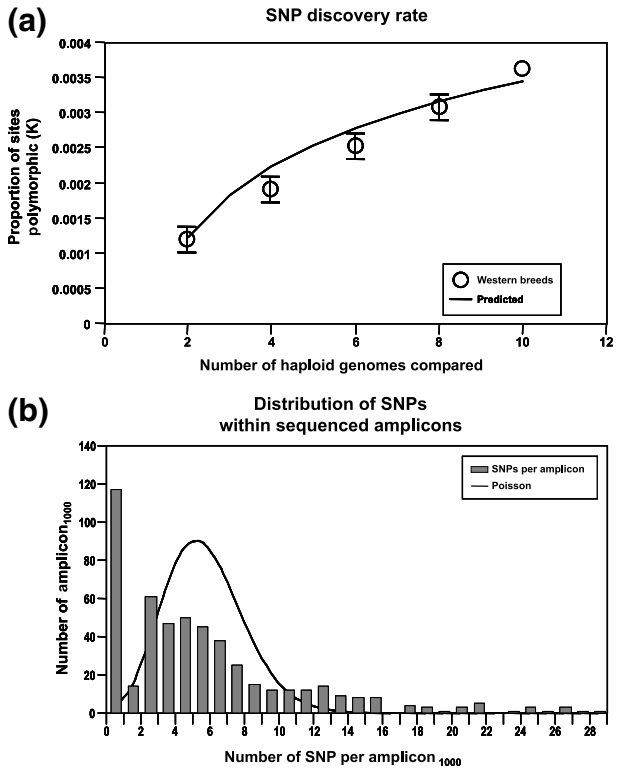

Figure 3 (a) Mean polymorphism frequencies detected given the inclusion of 1, 2, 3, 4 and 5 animals of western composition are presented (2, 4, 6, 8, and 10 haploid genomes). The SNP frequencies were predicted based upon the estimate that $K \approx \Theta [1^{-1} + 2^{-1} + 3^{-1} + \cdots + (n-1)^{-1}]$ and $H \approx \Theta$, where $H = 0.0012$ for comparison of two haploid genomes. (b) Distribution of porcine SNPs within PCR amplicons. Frequency of SNPs within each amplicon (SNP/bp) was determined. To enable comparison of amplicons of differing sizes, frequencies were converted to the value expected given 1000 bp of sequence surveyed. Distribution of SNP frequencies expected given adherence to the Poisson model and average SNP density is also present.

was made. Considering a heterogeneous set of amplicons with regard to type of primer-pair, predicted values for K ranged from 0.004 to 0.006. Thus, for MARCSDP a polymorphism can be found every 172–238 base pairs (average 184, Table 2). However, the likelihood that such predictions will be maintained at a genome-wide level in part depends on the adherence of frequency and distribution of porcine genetic diversity to that predicted by the infinite sites neutral theory model. To determine adherence of SNP yield to that predicted by the neutral model, five pigs of strictly western breed composition were assessed for K given all possible combinations and sample sizes. Despite the fact that all these animals are not from a single population in equilibrium, observed SNP frequencies closely approximate those predicted by the neutral model (Figs 3a; Table 4).

Although the number of SNPs observed met expectations predicted by the infinite sites neutral theory model, average density of SNPs within amplicons was not uniform (Fig. 3b). Number of amplicons without polymorphisms greatly exceeded that predicted by a Poisson sampling distribution, and SNP density in many amplicons surpassed predictions. Similar observations have also been made for mice and humans (Lindblad-Toh *et al.* 2000; Sachidanandam *et al.* 2001). This phenomenon may derive from either biochemical or evolutionary forces, such as regional differences in mutation and/or recombination rates, or differences of evolutionary selection experienced by different loci. Populations, which undergo intense selection pressure, such as livestock or domestic mice (Lindblad-Toh *et al.* 2000), or which have undergone dramatic bottlenecks in population size, are expected to be prone to uneven distribution in SNP density (Tajima 1983). While this effect does not seem to significantly impact total yield of SNPs during a discovery effort (Fig. 3a), it does influence SNP density within amplicons (Fig. 3b). Non-random distribution of SNPs in pigs resulted in failure to detect an SNP in 23% of amplicons generated. Efforts in identifying polymorphic loci in livestock should anticipate the influence of non-uniform SNP distribution on yield. As a result of non-uniform SNP distribution, refinement of QTL by LD mapping may vary dramatically among loci and between populations.

Conclusions

Our strategy for SNP discovery was based on the objective of refining the human/pig comparative genetic map using the MARC reference family. The subsequent mapping of these EST polymorphisms will substantially improve the current resolution of the comparative map. A total of 91 genes are on the current MARC swine linkage map (http://marc.usda.gov). The international collaborative effort (PIGMAP) has mapped by linkage 80 genes, 50 unique to those on the MARC map (http://www.thearkdb.org). Because a limited number of animals were sampled, utility of these identified polymorphisms for commercial applications has not yet been addressed. In fact, most of the polymorphisms identified were represented in Chinese × western breed contrast. While this reference family is extremely valuable for gene-mapping, not all sites that are polymorphic between European/western and Chinese breeds are of value for linkage disequilibrium mapping in commercial germplasm. In addition, SNPs that are limited to specific breeds are of special interest because of the possibility that they are of relatively recent origin and may contribute disproportionately to within-breed production trait variation, or will be in close association with causal polymorphisms. Utility of SNPs described here depends upon allele

Table 4 Summary of contributions by individual animals in the porcine SNP discovery panel for the amount of sequence obtained and the level of heterozygosity observed.

Breed composition[1]	Animal Number	Number amplicons	bp surveyed	Heterozygous amplicons	Number of bases heterozygous	Percentage heterozygous due to indel	H^2	bp between heterozygous positions
Meishan	199105908	336	133 977	82	160	6.9	0.0012	837
Meishan	199110803	320	119 246	84	143	9.8	0.0012	834
WC	199043006	488	220 549	120	217	10.6	0.0010	1016
WC	199045403	344	155 416	80	152	9.2	0.0010	1022
WC	199120807	342	151 501	98	198	7.1	0.0013	765
½ Duroc × ½ WC	199035602	460	206 640	128	221	7.6	0.0011	935
½ Duroc × ½ WC	199035802	481	220 655	130	258	7.8	0.0012	855
Minzhu	198893304	338	130 956	120	244	10.2	0.0019	537
7/8 Minzhu × 1/8 WC	199204909	351	158 057	125	254	5.9	0.0016	622
7/8 Minzhu × 1/8 WC	199205910	329	129 655	140	282	7.8	0.0022	460
½ Minzhu × ½ WC	199036703	473	191 383	196	390	7.4	0.0020	491
½ Minzhu × ½ WC	199036705	485	200 313	200	403	7.7	0.0020	497
½ Fengjing × ½ WC	199031304	478	201 991	206	445	7.9	0.0022	454
½ Meishan × ½ WC	199035702	348	135 644	139	284	7.7	0.0021	478
½ Meishan × ½ WC	199035902	484	194 464	205	415	8.2	0.0021	469
½ Meishan × ½ WC	199032102	498	205 595	223	429	7.0	0.0021	479

[1]WC = A composite population composed of ¼ Yorkshire, ¼ Large white, ¼ Chester white, ¼ Landrace from advanced generations of *inter se* matings.
[2]H = Heterozygosity per bp surveyed.

distribution and heterozygosity in breeds and herds important to commercial pork production. Therefore, determination of these values is critical to their application to industrial genetic improvement programmes.

Locus specific amplification has proven effective at identifying EST-associated polymorphisms for comparative map development. It has also provided an estimate of porcine sequence diversity, a parameter that will be useful for estimating expected rate of polymorphism discovery by genome-wide clone-based SNP discovery strategies. Clone-based discovery approaches do not require development of PCR primers, allow for immediate haplotype determination, and are not subject to sequence obfuscation by indels. As a result of efficiency of SNP discovery by RRS, it will probably provide a powerful and complementary approach to LSA for further SNP-based genetic map development in livestock.

Acknowledgements

Technical support was provided by Bree Quigley, Jeni Kuzsak, Renee Godtel, Bob Lee, Tracy Happold, Steve Simcox, Kevin Tennill, and Sam Nejezchleb. Mike Grosse and Eduardo Casas assisted with bovine primer testing. Jim Wray submitted the EST-associated STS to dbSTS. Sherry Kluver provided unparalleled secretarial support. Thanks to Dr Yang Da for many helpful discussions.

References

Altschul S.F., Wish W., Miller W., Myers E.W. & Lippman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–10.

Altshuler D., Pollard V.J., Cowles C.R., Van Etten W.J., Baldwin J., Linton L. & Lander E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407,** 513–6.

Buetow, K.H., Edmonson, M., MacDonald, R. *et al.* (2001) High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proceedings National Academy of Science USA* **98,** 581–4.

Chakravarti A. (1999) Population genetics-making sense out of sequence. *Nature Genetics* **21,** 56–60.

Deloukas, P., Schuler, G.D., Gyapay, G. *et al.* (1998) A physical map of 30,000 human genes. *Science* **282,** 744–6.

Ewing B. & Green P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8,** 186–94.

Ewing B., Hillier L., Wendl M.C. & Green P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8,** 175–85.

Fahrenkrug S.C., Smith T.P.L., Freking B.A. *et al.* (2002) Porcine gene-discovery by normalized cDNA-library sequencing and cluster assembly. *Mammalian Genome* (in press).

Gordon D., Abajian C. & Green P. (1998) CONSED: a graphical tool for sequence finishing. *Genome Research* **8,** 195–202.

Hartl D.L. & Clark A.G. (1997) *Principles of Population Genetics*, 3rd edn. Sinauer Associates, Inc, Publishers, Sunderlan, MA.

IHGSC (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

Krizman D.B. & Berget S.M. (1993) Efficient selection of 3′-terminal exons from vertebrate DNA. *Nucleic Acids Research* **21,** 5198–202.

Lindblad-Toh K., Winchester E., Daly M.J. *et al.* (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics* **24,** 381–6.

Long M., Rosenberg C. & Gilbert W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proceedings of the National Academy of Science USA* **92,** 12495–9.

Mullikin J.C., Hunt S.E., Cole C.G. *et al.* (2000) An SNP map of human chromosome 22. *Nature* **407,** 516–20.

Nickerson D.A., Tobe V.O. & Taylor S.L. (1997) POLYPHRED: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research* **25,** 2745–51.

Nickerson D.A., Taylor S.L., Weiss K.M., Clark A.G., Hutchinson R.G., Stengard J., Salomaa V., Vartiainen E., Boerwinkle E. & Sing C.F. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genetics* **19,** 233–40.

Rieder M.J., Taylor S.L., Tobe V.O. & Nickerson D.A. (1998) Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Research* **26,** 967–73.

Rohrer G.A., Alexander L.J., Keele J.W., Smith T.P. & Beattie C.W. (1994) A microsatellite linkage map of the porcine genome. *Genetics* **136,** 231–45.

Rozen S. & Skaletsky H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Molecular Biology* **132,** 365–86.

Sachidanandam R., Weissman D., Schmidt S.C. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–33.

Smith T.P.L., Grosse W.M., Freking B.A. *et al.* (2001a) Sequence evaluation of four pooled-tissue normalized bovine cDNA libraries and construction of a gene index for cattle. *Genome Research* **11,** 626–30.

Smith T.P.L., Fahrenkrug S.C., Rohrer G.A., Simmen F.A., Rexroad III C.E. & Keele J.W. (2001b) Mapping of expressed sequence tags from a porcine early embryonic cDNA library. *Animal Genetics* **32,** 66–72.

Tajima F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105,** 437–60.

Venter J.C., Adams M.D., Myers E.W. *et al.* (2001) The sequence of the human genome. *Science* **291,** 1304–51.

Wang D.G., Fan J.B., Siao C.J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–82.